Breakdown points for maximum likelihood-estimators of location-scale mixtures

Author(s):
Hennig, Christian

Publication date:
2002

Permanent link:
https://doi.org/10.3929/ethz-a-004336493

Originally published in:
Research Report / Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) 105

BREAKDOWN POINTS FOR MAXIMUM LIKELIHOOD-ESTIMATORS OF LOCATION-SCALE MIXTURES

by

Christian Hennig

Research Report No. 105
May 2002

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

BREAKDOWN POINTS FOR MAXIMUM LIKELIHOOD-ESTIMATORS OF LOCATION-SCALE MIXTURES

Christian Hennig

Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

May 2002

Abstract

ML-estimation based on mixtures of Normal distributions is a widely used tool for cluster analysis. However, a single outlier can break down the parameter estimation of at least one of the mixture components. Among others, the estimation of mixtures of t-distributions (McLachlan and Peel, 2000) and the addition of a further mixture component accounting for "noise" (Fraley and Raftery 1998) were suggested as more robust alternatives. In this paper, the definition of an adequate robustness measure for cluster analysis is discussed and bounds on the breakdown points of the mentioned methods are given. It turns out that the two alternatives, while adding stability in the presence of outliers of moderate size, do not possess a substantially better breakdown behavior than estimation based on Normal mixtures. If the number of clusters s is treated as fixed, r additional points suffice for all three methods to let the parameters of r clusters explode, unless $r = s$, where this is not possible for t-mixtures. The ability to estimate the number of mixture components, e.g., by use of the Bayesian Information Criterion (Schwarz 1978), and to isolate gross outliers as clusters of one point, is crucial for a better breakdown behavior of all three techniques. Furthermore, a sensible restriction of the parameter space to prevent singularities is discussed and a mixture of Normals with an improper uniform distribution is proposed for more robustness in the case of a fixed number of components.

Keywords: Model-based cluster analysis, robust statistics, mixtures of t-distributions, Normal mixtures, noise component, classification breakdown point

1 Introduction

ML-estimation based on mixtures of Normal distributions (NMML) is a flexible and widely used technique for cluster analysis (e.g., Fraley and Raftery 1998, Wang and Zhang 2002). Moreover, it is applied in density estimation and discrimination (Roeder and Wasserman 1997, Hastie and Tibshirani 1996). Banfield and Raftery (1993) introduced the term "model based cluster analysis" for such methods.

Observations $x_1, \ldots, x_n$ are modeled as i.i.d. according to the density

$$f_\eta(x) = \sum_{j=1}^{s} \pi_j \varphi_{a_j,\sigma_j^2}(x), \tag{1.1}$$

where $\eta = (s, a_1, \ldots, a_s, \sigma_1, \ldots, \sigma_s, \pi_1, \ldots, \pi_s)$ is the parameter vector, the number of components $s \in I\!N$ may be known or unknown, (a_j, σ_j) pairwise distinct, $a_j \in I\!R$, $\sigma_j > 0$, $\pi_j \geq$

$0,\ j = 1, \ldots, s$, $\sum_{j=1}^{s} \pi_j = 1$, and φ_{a,σ^2} denotes the density of a Normal distribution with mean a and variance σ^2, $\varphi = \varphi_{0,1}$. Often mixtures of multivariate Normals are used, but for the sake of simplicity, I restrict considerations to the case of one-dimensional data in this paper. The qualitative results should carry over to the multivariate case.

As many other ML-techniques based on the Normal distribution, NMML is not robust against gross outliers, at least if the number of components s is treated as fixed: The estimators of the parameters $a_1, \ldots, a_s$ are weighted means of the observations where the weights for each observation sum up to one, see (2.13), (2.19), which means that at least one of these parameters can get arbitrarily large if a single extreme point is added to a dataset.

There are some ideas to overcome the robustness problems of Normal mixture. The software `MCLUST` (Fraley and Raftery 1998) allows the addition of a mixture component accounting for "noise", modeled as a uniform distribution on the convex hull (the range in one dimension, respectively) of the data, and the software `EMMIX` (Peel and McLachlan 2000) can be used to fit a mixture of t-distributions instead of Normals. Further, it has been proposed to estimate the component parameters by more robust estimators (Campbell 1984, McLachlan and Basford 1988, Kharin 1996, p. 275), in particular by Huber's (1964, 1981) M-estimators corresponding to ML-estimation for a mixture of Huber's least favorable distributions (Huber 1964).

While a clear gain of stability can be demonstrated for these methods in various examples (see e.g. Banfield and Raftery 1993, McLachlan and Peel 2000, p. 231 ff.), there is a lack of theoretical justification of their robustness. Only Kharin (1996, p. 272 ff.) obtained some results for fixed s, showing that under certain assumptions on the speed of convergence of the proportion of contamination to 0 with $n \to \infty$, Huber's M-estimation is asymptotically superior to NMML. In this paper, mixtures of a class of location-scale models are considered, which includes the aforementioned distributions. The addition of a "noise"-component is also investigated.

In Section 2, the techniques treated in this paper and their underlying models are introduced. Some attention is paid to the restriction of the parameter space, which becomes necessary to define the ML-estimators properly, because the log-likelihood function of (1.1) and the other models can converge to ∞ if one of the σ_j^2 converges to 0. The restriction will usually have the form $\min_j \sigma_j \geq \sigma_0 > 0$. The choice of σ_0 has an impact to the stability properties of the methods. The alternative $\min_{j,k} \sigma_j/\sigma_k \geq c > 0$ (Hathaway 1985) is also discussed.

One of the problems is the difficulty to define an adequate measure of robustness for cluster analysis. In model based cluster analysis, the clusters are characterized by the parameters of their mixture components. For fixed s, an influence function (Hampel 1974) and a breakdown point (Hampel 1971, Donoho and Huber 1983) for these parameters can be defined straight forward. For the case of partitioning methods like k-means, k-medoids and trimmed k-means, where no variance is estimated, this has been done by Garcia-Escudero and Gordaliza (1999). They observed that a bounded influence function for these techniques does not necessarily imply a breakdown point larger than $1/(n+1)$, i.e., that it is not possible to let at least one of the parameters converge to ∞ by addition of a single point to the dataset. Note that an "addition" or "contamination" breakdown point is considered in this paper, which is computed by adding points to the original sample, while Garcia-Escudero and Gordaliza (1999) study "replacement breakdown". See Zhang and Li (1998), Zuo (2001) for relations between these two concepts. Zuo (2001) shows that under some assumptions (in particular dependence of the breakdown point on the data only through n) addition and replacement breakdown points are equivalent, but this does not hold in the setup considered here, as shown in Remark 4.17. It can further be distinguished between breakdown of a single cluster and breakdown of all clusters (Gallegos 2001).

Furthermore the breakdown point, for techniques where it does not always equal $1/(n+1)$, depends on the constellation of the data points, which lead to more or less stable clusterings. If the number of components s is estimated, the essentially unstable nature of cluster analysis becomes even clearer, because s is discrete and there must be data constellations "on the border" between two different numbers of components, leading to different numbers of parameters to estimate.

If the mixture model is applied for the aim of clustering, robustness could be defined in terms of the classification of the data points to the clusters as well. In this case it must be defined what change of a classification should be regarded as breakdown. Kharin (1996, p. 49) derives a decision rule for new points from estimators of mixture model parameters, and he defines breakdown as the degeneration of this rule to equiprobable coin-tossing. This, however, does not generalize straight forward to the case of estimated s.

In Section 3 I discuss robustness measures and breakdown points in terms of parameters as well as of classification, which are applicable for an estimated number of clusters. The definitions will be flexible enough to account for the breakdown of a single mixture component, for the breakdown of all mixture components and for intermediate situations. It is shown that breakdown of parameters and breakdown of classification do not always occur together.

In Section 4, some results about the parameter breakdown of the mixture based clustering techniques are derived. It is shown that all discussed techniques have a breakdown point of $r/(n+r)$ for $r < s$ of the mixture components in the case of fixed s. Only the breakdown of all s clusters by adding s points can be prevented for t-mixtures, as opposed to Normal mixtures. A better breakdown behavior can be attained by maximizing a kind of "improper likelihood" where "noise" is modeled by an improper uniform distribution on the real line. For the case of estimated s by use of an information criterion (Akaike 1974, Schwarz 1978), a breakdown point larger than $1/(n+1)$ can be reached for all treated methods. They all are able to isolate gross outliers as new mixture components on their own and are therefore very stable against extreme outliers. However, breakdown can happen because additional points inside the area of the estimated mixture components of the original data can lead to the estimation of a smaller number of components. The breakdown point depends on the constellation of the data in all cases. Some constellations are so stable that they have a breakdown point of larger than 1/2 to the price of a huge increase of the estimated number of clusters. The results can be interpreted as a characterization of the stability of the clustering of the concrete data. The paper is completed by some concluding discussions.

2 Models and methods

The Normal mixture (1.1) belongs to the class of mixtures of location-scale families which can be defined as follows:

$$f_\eta(x) = \sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x), \text{where } f_{a,\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x-a}{\sigma}\right). \tag{2.1}$$

η is defined as in (1.1). I assume that

$$f \text{ is symmetrical about } 0, \tag{2.2}$$

$$f \text{ decreases monotonously on } [0,\infty], \tag{2.3}$$

$$f > 0 \text{ on } \mathbb{R}, \tag{2.4}$$

$$f \text{ continuous.} \tag{2.5}$$

These assumptions are fulfilled, e.g., for the t_ν-distribution with ν degrees of freedom and for Huber's least favorable distribution, used as a basis for mixture modeling in Peel and McLachlan (2000), McLachlan and Basford (1988), respectively, besides the $\mathcal{N}(0,1)$-distribution.

Here are some consequences, which are needed later: It follows from (2.2)-(2.4) that for given points $x_1,\ldots,x_n$ and a compact set $C=[a,b]\times[\mu,\xi]\subset I\!R\times I\!R^+$ (this notation implies $\mu>0$ here)

$$\inf\{f_{a,\sigma}(x):\ x\in\{x_1,\ldots,x_n\},(a,\sigma)\in C\}=f_{min}>0. \tag{2.6}$$

Further, observe that for fixed x, $\lim_{m\to\infty}a_m=\infty$ and arbitrary sequences $(\sigma_m)_{m\in I\!N}$, as long as $\sigma_m\geq\sigma_0>0$,

$$\lim_{m\to\infty}f_{a_m,\sigma_m}(x)\leq\lim_{m\to\infty}\min\left(\frac{1}{\sigma_m}f(0),\frac{1}{\sigma_0}f\left(\frac{x-a_m}{\sigma_m}\right)\right)=0. \tag{2.7}$$

The addition of a uniform mixture component on the range of the data is also treated, which is the one-dimensional case of a suggestion of Banfield and Raftery (1993), that is, for given $x_{min},x_{max}\in I\!R$,

$$f_\zeta(x)=\sum_{j=1}^s\pi_j f_{a_j,\sigma_j}(x)+\pi_0\frac{1(x\in[x_{min},x_{max}])}{x_{max}-x_{min}}, \tag{2.8}$$

where $\zeta=(s,a_1,\ldots,a_s,\sigma_1,\ldots,\sigma_s,\pi_0,\pi_1,\ldots,\pi_s)$, $\pi_0,\ldots,\pi_s\geq 0$, $\sum_{j=0}^s\pi_j=1$ and $1(A)$ is the indicator function for the statement A.

I consider finite sample breakdown points as discussed in Donoho and Huber (1983). These are calculated from datasets $\mathbf{x}_n=(x_1,\ldots,x_n)$ and do not rest on a model assumption for the data. The presented mixture models are introduced to define the estimation procedures. For fixed s, parameters should be estimated by maximum likelihood (ML), where the data are treated as i.i.d. according to one of the models specified above (but ML estimation for different models may be applied to the same data). The log-likelihood functions for the models (2.1) and (2.8) at given data $\mathbf{x}_n$ with minimum $x_{min,n}$ and maximum $x_{max,n}$ (this notation is also used later) are

$$L_{n,s}(\eta,\mathbf{x}_n)=\sum_{i=1}^n\log\left(\sum_{j=1}^s\pi_j f_{a_j,\sigma_j}(x_i)\right), \tag{2.9}$$

$$L_{n,s}(\zeta,\mathbf{x}_n)=\sum_{i=1}^n\log\left(\sum_{j=1}^s\pi_j f_{a_j,\sigma_j}(x_i)+\frac{\pi_0}{x_{max,n}-x_{min,n}}\right). \tag{2.10}$$

$\mathbf{x}_n$ will be omitted if no confusion is possible. As can be seen easily by setting $a_1=x_1$, $\sigma_1\to 0$, $L_{n,s}$ is unbounded for $s>1$. That is, to define a proper ML estimator, the parameter space must be restricted somehow. The easiest restriction is to specify $\sigma_0>0$ and to demand

$$\sigma_j\geq\sigma_0,\ j=1,\ldots,s. \tag{2.11}$$

This is used, e.g., in DeSarbo and Cron (1988) and may easily be implemented in the EM-algorithm (Dempster, Laird and Rubin 1977, Redner and Walker 1984, see Lemma 2.1), the most frequently used routine to compute mixture ML estimators. The problem of this restriction is that the resulting ML estimators are no longer scale equivariant, because the scale of the data can be made arbitrarily small by multiplication with a constant. The alternative restriction

$$\min_{j,k=1,\ldots,s}\sigma_j/\sigma_k\geq c \tag{2.12}$$

for fixed $c \in (0, 1]$ leads to properly defined, scale equivariant, consistent ML estimators for the Normal case $f = \varphi_{0,1}$ without noise (Hathaway 1985). This includes the popular simplification $\sigma_1 = \ldots = \sigma_s$, which corresponds to k-means clustering and is the one-dimensional case of some of the covariance parameterizations implemented in `MCLUST` (Fraley and Raftery 1999). However, unless $c = 1$, the computation is not straightforward (Hathaway 1986). Furthermore, the restriction (2.12) cannot be applied to the model (2.8), because the log-likelihood function is not prevented from being unbounded, see Lemma 6.1. For the case of fixed s, Corollary 4.5 says that estimation using (2.12) does not own better breakdown properties than its counterpart using (2.11). Therefore, the restriction (2.11) is used unless indicated explicitly. Guidelines for the choice of σ_0 and c are given in Appendix 6.1. For results about consistency of local maximizers of the log-likelihood function see Redner and Walker (1984).

The following lemmas are used to establish certain properties and especially the existence of global maximizers of $L_{n,s}$ for both of the models (2.1) and (2.8). Notation: Let $\theta_j = (a_j, \sigma_j),\ j = 1, \ldots, s,\ \theta = (\theta_1, \ldots, \theta_s)$ denote the location and scale parameters of η, ζ, respectively, $\theta^*, \eta^*, \zeta^*$ by analogy (later the corresponding single parameters will be denoted by s^*, a_1^*, π_1^* and so on, and by analogy for $\hat{\eta},\ \hat{\zeta}\ldots$).

Lemma 2.1 *Let for given η*

$$p_{ij} = \frac{\pi_j f_{a_j,\sigma_j}(x_i)}{\sum_{k=1}^s \pi_k f_{a_k,\sigma_k}(x_i)},\ i = 1, \ldots, n. \tag{2.13}$$

A maximizer $\hat{\eta}$ of

$$\sum_{j=1}^s \left[\sum_{i=1}^n p_{ij} \log \pi_j^*\right] + \sum_{j=1}^s \sum_{i=1}^n p_{ij} \log f_{a_j^*,\sigma_j^*}(x_i) \tag{2.14}$$

over η^* leads to an improvement of $L_{n,s}$ unless η itself attains the maximum of (2.14).

For given ζ in (2.10) the same statements hold with

$$\begin{aligned} p_{ij} &= \frac{\pi_j f_{a_j,\sigma_j}(x_i)}{\sum_{k=1}^s \pi_k f_{a_k,\sigma_k}(x_i) + \pi_0/(x_{max,n} - x_{min,n})},\ j = 1, \ldots, s, \\ p_{i0} &= \frac{\pi_0/(x_{max,n} - x_{min,n})}{\sum_{k=1}^s \pi_k f_{a_k,\sigma_k}(x_i) + \pi_0/(x_{max,n} - x_{min,n})}. \end{aligned} \tag{2.15}$$

In (2.14), the first sum must start at $j = 0$.

This is derived in Redner and Walker (1984); for the results concerning ζ see DasGupta and Raftery (1998). (2.13) defines the so-called E-step and maximization of (2.14) defines the so-called M-step of the EM-algorithm, where the two steps are carried out alternately.

Lemma 2.2 *For any global maximizer η, ζ, respectively, of $L_{n,s}$ for given $\mathbf{x}_n$ under (2.11) the following conditions hold for $j = 1, \ldots, s$ with $p_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, s$ as in (2.15):*

$$\pi_j = \tfrac{1}{n} \sum_{i=1}^n p_{ij}, \tag{2.16}$$

$$(a_j, \sigma_j) = \arg\max S_j(a_j^*, \sigma_j^*) = \arg\max \sum_{i=1}^n p_{ij} \log\left(\frac{1}{\sigma_j^*} f\left(\frac{x_i - a_j^*}{\sigma_j^*}\right)\right). \tag{2.17}$$

In the case of (2.10), (2.16) holds as well for $j = 0$.

With $C = [x_{min,n}, x_{max,n}] \times \left[\sigma_0, \frac{\sigma_0 f(0)}{f\left(\frac{x_{max,n} - x_{min,n}}{\sigma_0}\right)}\right]$ and $\pi_1, \ldots, \pi_s > 0$,

$$\forall \theta^* \notin C^s\ \exists \theta \in C^s:\ L_{n,s}(\eta) > L_{n,s}(\eta^*). \tag{2.18}$$

All proofs are given in the Appendix 6.2.

For Normal f, the solutions of (2.17) for $j = 1, \ldots, s$ are

$$a_j = \frac{\sum_{i=1}^n p_{ij} x_i}{\sum_{i=1}^n p_{ij}},\ \sigma_j^2 = \frac{\sum_{i=1}^n p_{ij}(x_i - a_j)^2}{\sum_{i=1}^n p_{ij}}, \tag{2.19}$$

as long as this does not lead to $\sigma_j < \sigma_0$, see Redner and Walker (1984).

$L_{n,s}$ is continuous because of (2.5) and a global maximizer must lie in $C^s \times [0,1]^s$ because of (2.18). Therefore

Corollary 2.3 *Under the restriction (2.11) there exists a (not necessarily unique) global maximum of $L_{n,s}$ with arguments in $C^s \times [0,1]^s$.*

For NMML and (2.12), this is shown by Hathaway (1985). Define $\eta_{n,s} = \arg\max L_{n,s}$, analogously $\zeta_{n,s}$. In the case of non-uniqueness, $\eta_{n,s}$ can be defined as an arbitrary maximizer, e.g., the lexicographically smallest one. The p_{ij}-values from (2.13), (2.15), respectively, can be interpreted as the a posteriori probabilities that a point x_i had been generated by component j under the a priori probability π_j for component j with parameters a_j, σ_j. These values can be used to classify the points and to generate a clustering by

$$l(x_i) = \arg\max_j p_{ij},\ i = 1, \ldots, n, \tag{2.20}$$

where the ML-estimator is plugged into the definition of p_{ij}.

For the breakdown considerations in Section 4 it is necessary to investigate if $\eta_{n+g,s}$ stays inside some compact set or leaves it under the addition of g points. All theorems will hold for any of the maximizers. For ease of notation, $\eta_{n,s}$ and $\zeta_{n,s}$ will be treated as well defined in the following. Note that for $s > 1$ non-uniqueness occurs at least because of "label switching" of the mixture components. Further, for the ease of notation, it is not assumed that $\pi_j > 0\ \forall j$ or all (a_j, σ_j) being pairwise distinct.

Consider now the number of mixture components $s \in I\!N$ as unknown. The most popular method to estimate s is the use of information based criteria such as AIC (Akaike 1974) and BIC (Schwarz 1978). The latter is implemented in `MCLUST`. `EMMIX` computes them both. The estimator s_n for the correct order of the model is defined as $s_n = \arg\max_s C(s)$, where

$$\begin{aligned} C(s) = \mathrm{AIC}(s) &= 2L_{n,s}(\eta_{n,s}) - 2k, \\ C(s) = \mathrm{BIC}(s) &= 2L_{n,s}(\eta_{n,s}) - k \log n, \end{aligned} \tag{2.21}$$

respectively, where k denotes the number of free parameters, i.e., $k = 3s - 1$ for (2.1) and $k = 3s$ for (2.8). Under assumptions which hold for the models discussed here under (2.11) but not under (2.12) (compare Lemma 6.1), Lindsay (1995, p.22) shows that the number of distinct points in the dataset is an upper bound for the maximization of $L_{n,s}(\eta_{n,s})$ over s, and therefore as well for the maximization of $C(s)$. Therefore only finitely many values for s have to be checked to maximize $C(s)$ and this means that (again non-unique) maximizers exist.

While the AIC is known to overestimate s asymptotically (see, e.g., Bozdogan 1994), the BIC is shown at least in some restricted situations to be consistent in the mixture setup (Keribin 2000). I mainly consider the BIC here. Further suggestions to estimate s, which are more difficult to analyze with respect to the breakdown properties, are given, e.g., by Bozdogan (1994) and Celeux and Soromenho (1996). `EMMIX` also allows the estimation of s via a bootstrapped likelihood ratio test (McLachlan 1987).

3 Breakdown measures for cluster analysis

The classical meaning of breakdown for finite samples is that an estimator can be driven as far away from its original value as possible by addition of arbitrarily unfortunate points, usually by gross outliers. This holds for the here considered "addition breakdown points" as opposed to "replacement breakdown points", where points from the original sample are replaced (Donoho and Huber 1983, Zhang and Li 1998). Breakdown means that estimators, which can take values on the whole range of $\mathbb{R}^p$, can leave every compact set. If the value range of a parameter is bounded, breakdown means that addition of points can take the parameter arbitrarily close to the bound, e.g., a scale parameter to 0. Such a definition can be applied relatively easily to the estimation of mixture components, but it cannot be used to compare the robustness of mixture estimators with other methods of cluster analysis.

Therefore I will propose a breakdown definition in terms of the classification of points to clusters after the definition of the more familiar parameter breakdown point.

A "parameter breakdown" can be understood in two ways: A situation where at least one of the mixture components explodes is defined as breakdown in Garcia-Escudero and Gordaliza (1999). That is, breakdown occurs if the whole parameter vector leaves all compact sets (not including scales of 0 under (2.12)). In contrast to that, Gallegos (2001) defines breakdown in cluster analysis as a situation where *all* clusters explode simultaneously. Intermediate situations may be of interest in practice, especially if a researcher is interested in preventing breakdown of a single cluster by specifying the number of clusters larger than expected to catch the outliers. This is discussed (but not recommended - in agreement with the results given here) by Peel and McLachlan (2000). The definition given here is flexible enough to account for all these situations.

Definition 3.1 *Let $(E_n)_{n\in\mathbb{N}}$ be a sequence of estimators of η in model (2.1), of ζ in model (2.8), respectively, on $\mathbb{R}^n$ for fixed $s \in \mathbb{N}$. Let $r \leq s$, $\mathbf{x}_n = (x_1, \ldots, x_n)$ be a dataset, where*

$$\forall \hat{\eta} = \arg\max_{\eta} L_{n,s}(\eta, \mathbf{x}_n): \ \hat{\pi}_j > 0, \ j = 1, \ldots, s. \tag{3.1}$$

The r**-components parameter breakdown point** *of E_n is defined as*

$$\begin{gathered} B_{r,n}(E_n, \mathbf{x}_n) = \min\nolimits_g \{ \tfrac{g}{n+g} : \ \exists j_1 < \ldots < j_r \\ \forall \ D = [\pi_{min}, 1] \times C, \ \pi_{min} > 0, \ C \subset \mathbb{R} \times \mathbb{R}^+ \text{ compact} \\ \exists \ \mathbf{x}_{n+g} = (x_1, \ldots, x_{n+g}), \ \hat{\eta} = E_{n+g}(\mathbf{x}_{n+g}) : \ (\hat{\pi}_j, \hat{a}_j, \hat{\sigma}_j) \notin D, \ j = j_1, \ldots, j_r \}. \end{gathered}$$

The proportions π_j are defined not to break down if they are bounded away from 0, which implies that they are bounded away from 1 if $s > 1$. Assumption (3.1) is necessary for the definition to make sense; $\hat{\pi}_j = 0$ would imply that the corresponding location and scale parameters could be chosen arbitrarily far out without adding any point. (3.1) may be violated in situations where s is large compared to n, but these situations are usually not of interest in cluster analysis. In particular, (3.1) does not hold if s exceeds the number of distinct x_i, see Lindsay (1995, p. 23).

$\hat{\pi}_0 \to 0$ in model (2.8) is not defined as breakdown, because the noise component is not considered as an object of interest in itself in this setup.

In the situation for unknown s, I restrict considerations to the case of 1-components breakdown, because this enables already a satisfying breakdown behavior of the usual mixture methods. Breakdown means that neither of the s mixture components estimated for $\mathbf{x}_n$ vanishes, nor that any of their scale and location parameters explodes to ∞ under addition of points. It is however allowed that the new dataset yields more than s mixture components, and that the additional mixture components have arbitrary parameters. This means that if the outliers form a cluster for themselves, their component can simply be added without breakdown. Further, breakdown of the proportions π_j to 0 is no longer of interest for estimated s according to the AIC or BIC, because if some π_j is small enough, component j can be simply left out, and the other proportions can be updated to sum up to 1. This solution with $s-1$ clusters leads approximately to the same log-likelihood and will be preferred because of the penalty on the number of components:

Definition 3.2 *Let* $(E_n)_{n\in\mathbb{N}}$ *be a sequence of estimators of* η *in model (2.1) or of* ζ *in model (2.8), respectively, on* $\mathbb{R}^n$*, where* $s \in \mathbb{N}$ *is unknown and estimated as well. Let* $\mathbf{x}_n = (x_1, \ldots, x_n)$ *be a dataset. Let* s *be the estimated number of components of* $E_n(\mathbf{x}_n)$*. The* **parameter breakdown point** *of* E_n *is defined as*

$$\begin{gathered}B_n(E_n, \mathbf{x}_n) = \min{}_g\{\tfrac{g}{n+g} : \ \forall C \subset \mathbb{R}^s \times (R^+)^s \text{ compact} \\ \exists\ \mathbf{x}_{n+g} = (x_1, \ldots, x_{n+g}),\ \hat\eta = E_{n+g}(\mathbf{x}_{n+g}) : \\ \text{pairwise distinct } j_1, \ldots, j_s \text{ do not exist, such that } (\hat a_{j_1}, \ldots, \hat a_{j_s}, \hat\sigma_{j_1}, \ldots, \hat\sigma_{j_s}) \in C\}.\end{gathered}$$

This implies especially that breakdown occurs whenever $\hat s < s$.

Now I turn to the definition of classification breakdown. A mapping E_n is called a general clustering method (GCM), if it maps a set of entities $\mathbf{x}_n = \{x_1, \ldots, x_n\}$ to a collection of its subsets $\{C_1, \ldots, C_s\}$. A special case are partitioning methods where $C_i \cap C_j = \emptyset$ for $i \neq j \leq s$, $\bigcup_{j=1}^{s} C_s = \mathbf{x}_n$. An ML-mixture estimator induces a partition by (2.20) and $C_j = \{x_i : \ l(x_i) = j\}$, given a rule to break ties in the p_{ij}.

If E_n is a GCM and $\mathbf{x}_{n+g}$ is generated by adding g points to $\mathbf{x}_n$, $E_{n+g}(\mathbf{x}_{n+g})$ induces a clustering on $\mathbf{x}_n$, which is denoted by $E_n^*(\mathbf{x}_{n+g})$. Its clusters are denoted by $C_1^*, \ldots, C_{s^*}^*$. If E_n is a partitioning method, $E_n^*(\mathbf{x}_{n+g})$ is a partition as well. s^* may be smaller than s when E_n produces s clusters for all n.

As will be illustrated in Section 4, different clusters of the same data may have a different stability for GCMs. Thus, I define robustness with respect to the single clusters. Therefore, a measure is needed for the similarity between some cluster of $E_n^*(\mathbf{x}_{n+g})$ and a cluster of $E_n(\mathbf{x}_n)$, i.e., between two subsets C and D of some finite set of entities. From lots of possibilities, I have chosen the following, which gives 0 only for disjoint sets and 1 only for equal sets:

$$\gamma(C, D) = \frac{2|C \cap D|}{|C| + |D|}.$$

The definition of breakdown bases on the similarity of a cluster from $E_n(\mathbf{x}_n)$ to its most similar cluster in $E_n^*(\mathbf{x}_{n+g})$. For $C \in E_n(\mathbf{x}_n)$ and an arbitrary GCM $\hat E_n$:

$$\gamma^*(C, \hat E_n(\mathbf{x}_n)) = \min_{D \in \hat E_n(\mathbf{x}_n)} \gamma(C, D).$$

How small should γ^* be to say that breakdown of C has occurred? The proposed answer is "$\leq \frac{2}{3}$". The reason is as follows: Suppose that a dataset of (even) n points is partitioned into 2 clusters C_1, C_2 of $\frac{n}{2}$ points. Suppose further, that by addition of g points, all n original points fall into the same cluster D. This means, for $j = 1, 2$,

$$\gamma^*(C_j, E_n^*(\mathbf{x}_{n+g})) = \frac{n}{n/2 + n} = \frac{2}{3}.$$

Further, as long as E_n is a partitioning method, observe

$$\exists C \in E_n(\mathbf{x}_n) : \gamma^*(C_j, E_n^*(\mathbf{x}_{n+g})) \leq \frac{2}{3}$$

whenever $|E_n(\mathbf{x}_n)| = s$ and $|E_n^*(\mathbf{x}_{n+g})| = s-1$, because in this case there must be $D \in E_n^*(\mathbf{x}_{n+g})$ such that there are at least two members of $E_n(\mathbf{x}_n)$, C_1 and C_2, say, for which D minimizes $\gamma(C_j, D)$ over $E_n^*(\mathbf{x}_{n+g})$. W.l.o.g., $|C_1 \cap D| \leq |C_2 \cap D|$. Because of $C_1 \cap C_2 = \emptyset$, get $\gamma(C_1, D) \leq \frac{|D|}{|D|/2+|D|}$. This means that at least one of the original clusters is said to break down if a cluster is lost in the induced partition, and $\frac{2}{3}$ is the smallest cutoff value for this to hold. Note further that at least $r > 1$ clusters must break down if $|E_n^*(\mathbf{x}_{n+g})| = s - r$, because the same arguments show that $\gamma(C_j, D) \leq \frac{2}{3}$ for at least $q-1$ clusters C_j, if $D \in E_n^*(\mathbf{x}_{n+g})$ is the most similar cluster for q clusters $C_j \in E_n(\mathbf{x}_n)$.

Definition 3.3 *Let $(E_n)_{n\in\mathbb{N}}$ be a sequence of GCMs. The* **classification breakdown point** *of a cluster $C \in E_n(\mathbf{x}_n)$ is defined as*

$$B_n^c(E_n, \mathbf{x}_n, C) = \min_g \left\{ \frac{g}{n+g} : \ \exists \mathbf{x}_{n+g} = (x_1, \ldots, x_{n+g}) : \ \gamma^*(C, E_n^*(\mathbf{x}_{n+g})) \leq \frac{2}{3} \right\}.$$

Because the cutoff value $\frac{2}{3}$ has been justified only for partitioning methods, it may be doubted that this definition should also be used for other GCMs.

Note that neither does parameter breakdown imply classification breakdown, nor the other way round, see the Remarks 4.12 and 4.20. Note further that the classification breakdown point differs from the classical definitions of breakdown points, because $\frac{2}{3}$ is not the worst possible value for $\gamma^*(C, E_n^*(\mathbf{x}_{n+g}))$. However, the worst possible value depends on the clustering of the original data and is therefore not suitable for a breakdown definition.

The classification breakdown point is more difficult to handle mathematically than the parameter breakdown point, but it is shown in Section 4 that it can be worked out sometimes.

4 Breakdown results

4.1 Breakdown points for fixed s

The section starts with three lemmas, which characterize the behavior of the estimators under addition of some points $x_{n+1}, \ldots, x_{n+g}$ in large enough distance to data $\mathbf{x}_n = (x_1, \ldots, x_n)$. In this case, the $s > 1$ mixture components estimated by ML can be separated into components corresponding to $\mathbf{x}_n$ and components corresponding to $(x_{n+1}, \ldots, x_{n+g})$. There exists at least one component corresponding to each of these two classes, and the maximum of the log-likelihood can be obtained from the maxima considering $\mathbf{x}_n$ and the outliers alone. This means in particular, that at least one of the s components of the original dataset $\mathbf{x}_n$ must break down under addition of a large enough gross outlier, and this does not depend on the basic distribution f used for the definition of the ML estimator. Further, if the $(x_{n+1}, \ldots, x_{n+g})$ can be divided into groups, the distances between which converge to ∞ as well, a mixture component can eventually be assigned to each of the groups as long as enough components exist (note that $n+g$ corresponds to n in the notation of the lemmas, and n corresponds to n_1). This is different from the case $s = 1$, where the parameters break down because of a single gross error under ML estimation for the Normal, but not for t_ν-distributions (Tyler 1994). A similar phenomenon occurs for k-medoids clustering, which is based on the robust median for one-dimensional data, but breaks down for $k > 1$ clusters (Garcia-Escudero and Gordaliza 1999). The lemmas are shown for $h \geq 2$ groups of points between which the distance converges to infinity. This means that at last all groups have to be fitted separately.

Lemma 4.1 *Let* $\mathbf{x}_{nm} = (x_{1m}, \ldots, x_{nm}) \in I\!R^n$, $m \in I\!N$, *where* $0 = n_0 < n_1 < \ldots < n_h = n$, $h \geq 1$, *be a sequence of datasets. Let* $D_1 = \{1, \ldots, n_1\}$, $D_2 = \{n_1 + 1, \ldots, n_2\}, \ldots, D_h = \{n_{h-1} + 1, \ldots, n_h\}$. *Assume further that*

$$\exists\, b < \infty : \max_k \max_{i,j \in D_k} |x_{im} - x_{jm}| \leq b \ \forall m,$$
$$\lim_{m \to \infty} \min_{k \neq l, i \in D_k, j \in D_l} |x_{im} - x_{jm}| = \infty.$$

Let $s \geq h$ *be fixed,* $\eta_m = \arg\max_\eta L_{n,s}(\eta, \mathbf{x}_{nm})$ *(parameters called* $\pi_{1m}, \ldots, \pi_{sm}, a_{1m}$ *and so on; all results hold for* ζ_m *from maximization of (2.10) as well). W.l.o.g.,* $x_{1m} \leq x_{2m} \leq \ldots \leq x_{nm}$. *Then, for large enough* $m_0 \in I\!N$,

$$\begin{aligned} &\exists 0 \leq d < \infty : \ \forall m > m_0, \ k = 1, \ldots, h \ \exists j_k \in \{1, \ldots, s\}, \ \pi_{min} > 0, \ \sigma_0 \leq \sigma_{max} < \infty : \\ &\quad a_{j_k m} \in [x_{(n_{k-1}+1)m} - d, x_{n_k m} + d], \ \pi_{j_k m} \geq \pi_{min}, \ \sigma_{j_k m} \in [\sigma_0, \sigma_{max}]. \end{aligned} \tag{4.1}$$

Lemma 4.2 *In the situation of Lemma 4.1, assume further*

$$\exists \pi_{min} > 0 : \ \forall j = 1, \ldots, s, \ m \in I\!N : \ \pi_{jm} \geq \pi_{min}. \tag{4.2}$$

Then,

$$\forall m > m_0, \ j = 1, \ldots, s \ \exists k \in \{1, \ldots, h\} : \ a_{jm} \in [x_{(n_{k-1}+1)m} - d, x_{n_k m} + d]. \tag{4.3}$$
$$\exists 0 \leq \sigma_{max} < \infty : \ \forall m > m_0, \ j = 1, \ldots, s : \ \sigma_{jm} \in [\sigma_0, \sigma_{max}]. \tag{4.4}$$

Lemma 4.3 *Under the assumptions of Lemma 4.1,*

$$\forall k \in \{1, \ldots, h\} : \ \lim_{m \to \infty} \sum_{a_{jm} \in [x_{(n_{k-1}+1)m} - d, x_{n_k m} + d]} \pi_{jm} = \frac{|D_k|}{n}, \tag{4.5}$$

$$\lim_{m \to \infty} \left| L_{n,s}(\eta_m, \mathbf{x}_{nm}) - \max_{\sum_{k=1}^h q_k = s} \left(\sum_{k=1}^h \left[\max_\eta L_{|D_k|, q_k}(\eta, \mathbf{y}_{km}) + |D_k| \log \frac{|D_k|}{n} \right] \right) \right| = 0, \tag{4.6}$$
$$\text{where } \mathbf{y}_{km} = (x_{(n_{k-1}+1)m}, \ldots, x_{n_k m}), \ k = 1, \ldots, h.$$

This means in particular, that $r < s$ added outliers, the difference between which goes to ∞, let r mixture components break down.

Theorem 4.4 *Let* $\mathbf{x}_n \in I\!R^n$, $s > 1$. *Let* $\eta_{n,s}$ *be a global maximizer of (2.9). Assume (2.2)-(2.5). For* $r = 1, \ldots, s - 1$,

$$B_{r,n}(\eta_{n,s}, \mathbf{x}_n) \leq \frac{r}{n + r}. \tag{4.7}$$

"=" in (4.7) could be proven for datasets where $\pi_j \to 0$ can be prevented for $j = 1, \ldots, s$ and any sequence of sets of r added points, but conditions for this are hard to derive.

Under the restriction (2.12), convergence of σ_j-parameters to 0 means breakdown according to Definition 3.1. Thus, to prevent breakdown, an effective lower bound for the σ_j of the non-breaking components must exist, and this means that all σ_j must be bounded from below, independently of $x_{n+1}, \ldots, x_{n+r}$, because (2.12) forces all σ_j to 0, if only one implodes. Therefore the result carries over:

Corollary 4.5 *Theorem 4.4 holds as well under the restriction (2.12) instead of (2.11).*

From now on, only (2.11) will be used.

The situation for $r = s$ is a bit more complicated. The choice of the basic distribution f does not matter until here. $s - 1$ outliers can break down $s - 1$ mixture components, but the sth mixture component must be a compromise between the original dataset $\mathbf{x}_n$ and the sth outlier. If such a compromise is estimated by a non-robust ML-estimator such as the Normal one, the sth component will break down as well.

Theorem 4.6 *Let* $\mathbf{x}_n \in I\!R^n$, $f = \varphi$. *Then,*

$$B_{s,n}(\eta_{n,s}, \mathbf{x}_n) \leq \frac{s}{n+s}.$$

The breakdown point for the joint ML-estimator of location and scale for a single location-scale model based on the t_ν-distribution was derived by Tyler (1994). Ignoring the possible case that the scale breaks down to 0, which is treated by Tyler but which is not of interest in the setup here because of (2.11), the breakdown point is shown to be $\geq \frac{1}{\nu+1}$ (equality is only distorted because of rounding to ratios of integers appearing in the definition of the breakdown point). This carries over to the mixture ML-estimator only for the breakdown of the last mixture component.

Theorem 4.7 *Let* $\mathbf{x}_n \in I\!R^n$, $f(x) = b\left(1 + \frac{x^2}{\nu}\right)^{(-\nu+1)/2}$, $\nu \geq 1$, $b > 0$ *being the appropriate norming constant. Then,*

$$B_{s,n}(\eta_{n,s}, \mathbf{x}_n) \geq \frac{1}{\nu+1}.$$

However, the t_ν-approach must be judged as essentially not breakdown-robust as long as s is fixed because of Theorem 4.4, even if it has an advantage over the Normal estimator.

Unfortunately, the approach via adding a noise component does not lead to better breakdown results, because a single outlier can make the density value of the noise component arbitrarily small, so that again solutions with one-point mixture components for the outliers are better in terms of the log-likelihood than it would be to classify them as noise.

Theorem 4.8 *Theorem 4.4 and Theorem 4.6 hold as well for global maximizers of (2.10).*

Example 4.9 *While the breakdown point for all treated approaches is the same for* $r < s$, *it may be of interest, how large an outlier must be to cause breakdown of the methods. The following definition is used to generate reproducible example datasets:*

Definition 4.10 $\Phi^{-1}_{a,\sigma^2}(\frac{1}{n+1}), \ldots, \Phi^{-1}_{a,\sigma^2}(\frac{n}{n+1})$ *is called a* (a, σ^2)-**Normal standard dataset** *(NSD) with n points, where* Φ_{a,σ^2} *denotes the cdf of the Normal distribution with parameters* a, σ^2.

Consider a dataset of 50 points, namely a (0,1)-NSD with 25 points combined with a (5,1)-NSD with 25 points, see Figure 1, and $s = 2$. *For Normal mixtures,* t_μ*-mixtures with* $\mu > 1$ *and Normal mixtures with noise component, always components corresponding almost exactly to the two NSDs are optimal under* $\sigma_0 = 0.025$ *(see Example 4.14 for the rationale behind this choice). How large must an additional outlier be chosen so that the 50 original points*

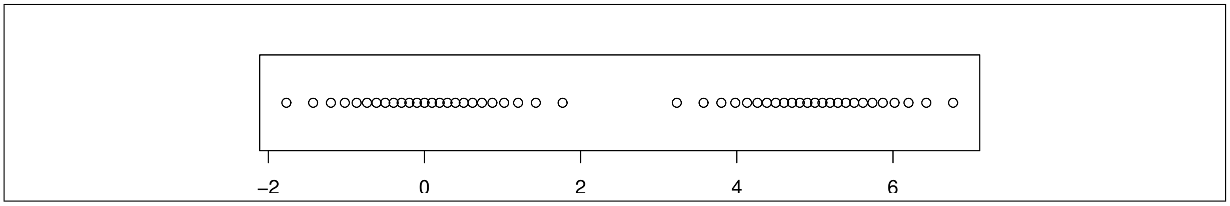

Figure 1: “Standard” example dataset: 25 points (0,1)-NSD combined with 25 points (5,1)-NSD.

fall into only one cluster and the second mixture component fits only the outlier? This makes a strong distinction between the estimators. For Normal mixtures, breakdown begins with an additional point at about 15.2. For a mixture of t_3-distributions, the outlier must lie at about 800, t_1-mixtures need the outlier at about $3.8*10^7$, and a Normal mixture with additional noise component breaks down with an additional point at $3.5*10^7$. These values, however, depend strongly upon σ_0.

An initial classification of some points as noise is needed to fit NMML with noise component in `MCLUST`. This can be done for example by use of the nearest neighbor clutter removal of Byers and Raftery (1998). Theorem 4.8 holds for the global optimum of (2.10) and one may wonder whether the `MCLUST`-fit might be more robust when extreme outliers are initially classified as noise (which happens under nearest neighbor clutter removal). But this allows at most one point more before breakdown occurs.

Lemma 4.11 *Let $\mathbf{x}_n$ be a fixed dataset. Let η_{EM} be defined as an arbitrary limit point of the EM algorithm locally optimizing (2.10) in the case $f = \varphi$ such that the initial values for datasets consisting of $\mathbf{x}_n$ and some further points $x_{n+1}, \ldots, x_{n+g}$ satisfy*

$$i > n \Rightarrow p_{i0} = 1,\ \forall j \in \{1, \ldots, s\}\ \exists i \in \{1, \ldots, n\} : p_{ij} > 0. \tag{4.8}$$

Assume that for all steps of the EM-algorithm the scale parameters of the maximizers of (2.17) are $\geq \sigma_0$ for unrestricted σ. Then

$$B_{1,n}(\eta_{EM}, \mathbf{x}_n) \leq \frac{2}{n+2}. \tag{4.9}$$

Assume further that some $\pi_{min} > 0$ exists such that $\pi_j \geq \pi_{min}$ for fixed j and all steps of the EM-algorithm. If $g = 1$ and $|x_{n+1}|$ large enough, $\epsilon > 0$ arbitrary,

$$(a_{EM,j}, \sigma_{EM,j}) \in C_\epsilon,\ j = 1, \ldots, s, \tag{4.10}$$

where $C_\epsilon = [x_{min,n} - \epsilon, x_{max,n} + \epsilon] \times [\sigma_0, \sqrt{(x_{max,n} - x_{min,n})^2 + \epsilon}]$.

This means that breakdown with only one additional point is only possible because of some π_j converging to 0, which can only happen under very unstable data constellations $\mathbf{x}_n$.

Note further that x_{n+2} must be extremely huge to cause a moderately large x_{n+1} to be merged with one of the Normal mixture components. The exponentially decreasing density value of x_{n+1} under one of the Normal components arising from $x_1, \ldots, x_n$ must get larger than $\frac{1}{x_{n+2} - x_{min,n}}$. For example, x_{n+2} must be chosen between 10^{110} and 10^{120} to merge $x_{n+1} = 20$ with an (0,1)-NSD of size 20 when $s = 1$ component plus noise is to be fitted (σ_0 chosen large enough that the non-noise component corresponds to the whole NSD). This behavior does not change when the number of outliers is enlarged. Only the most extreme point matters. Therefore `MCLUST` together with a good initial noise detector can be expected to be relatively stable in practice as long as disastrous outliers are removed beforehand.

Remark 4.12 *Theorems 4.4 and 4.8 carry over to the classification breakdown point in the sense that under the given circumstances at least* r *of the original clusters must break down. This follows because if* r *outliers are added, converging to* ∞ *and with the distance between them converging to* ∞ *as well, Lemma 4.2 yields that* $p_{ij} \to 0$ *for the original points* $i = 1, \ldots, n$ *and* j *fulfilling* $a_{jm} \in [x_{n+g} - d, x_{n+g} + d]$ *for some* $g \in \{1, \ldots, r\}$*. That is, at most* $s - r$ *clusters remain for the classification of the original points. On the other hand, the arguments leading to Theorem 4.6 do not carry over, because the addition of* $r = s$ *outliers as above certainly explodes all mean parameters, but one cluster usually remains that contains all the original points. Therefore, an original cluster containing more than half of the points does not break down in the sense of classification.*

4.2 Alternatives for fixed s

The results given previously indicate that the treated mixture methods are generally not robust against breakdown for fixed s. There are two principles which may lead to a better breakdown behavior. The first principle is to optimize a target function for only a part of the data, say, optimally selected 50% or 80% of the points. The methods of trimmed k-means (Garcia-Escudero and Gordaliza 1999) and clustering based on minimum covariance determinant estimators (Rocke and Woodruff 2000, Gallegos 2001) use this principle. Both methods, however, rest on a partition model as opposed to the mixture model. This may be useful for clustering, but leads to biased parameter estimators (Bryant and Williamson 1986).

Another alternative can be constructed as a modification of the uniform noise approach. The problem of this approach is that the noise component could be affected by outliers as well, as was shown in the previous section. This can be prevented when the density constant for the noise component is chosen as fixed beforehand, which leads to ML estimation for a mixture where some improper distribution component is added to catch the noise. That is, an estimator $\xi_{n,s}$ is defined as the maximizer of

$$L_{n,s}(\xi, \mathbf{x}_n) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{s} \pi_j f_{a_j,\sigma_j}(x_i) + \pi_0 b \right), \tag{4.11}$$

where $b > 0$. This requires the choice of b. If the objective is cluster analysis and there is a maximum scale σ_{max}, above which a mixture component is no longer accepted as a cluster (compare Appendix 6.1), b could be chosen as the density value at the 0.025-quantile of $f_{0,\sigma_{max}}$, so that 95% of the points generated from such a distribution have a larger density value for it than for the noise component. For this estimator the breakdown point depends on the stability of the dataset $\mathbf{x}_n$. Breakdown can only occur if additional observations allow that the non-outliers can be fitted with advantage by fewer than s components, and this means that a relatively good solution for $r < s$ components must exist already for $\mathbf{x}_n$. This is formalized in (4.12). Let $L_{n,s} = L_{n,s}(\xi_{n,s}, \mathbf{x}_n)$. I consider only the breakdown of a single mixture component $B_{1,n}(\xi_{n,s}, \mathbf{x}_n)$. Breakdown points for more than one component must be larger.

Theorem 4.13 *Let* $\mathbf{x}_n \in I\!R^n$. *Let* $\xi = \xi_{n,s}$ *and* $f_{max} = f(0)/\sigma_0 > b$. *If*

$$\begin{aligned}\max_{r<s} L_{n,r} \quad < \quad & \sum_{i=1}^{n} \log\left(\sum_{j=1}^{s} \pi_j f_{\theta_j}(x_i) + (\pi_0 + \frac{g}{n})b\right) \\ & + g\log(\pi_0 + \frac{g}{n})b + (n+g)\log\frac{n}{n+g} - g\log f_{max}, \qquad (4.12)\end{aligned}$$

then

$$B_{1,n}(\xi_{n,s}, \mathbf{x}_n) \geq \frac{g}{n+g}. \qquad (4.13)$$

Example 4.14 *Here are three examples to illustrate the meaning of (4.12) in case of* $f = \varphi$.

Consider first the dataset of 50 points which is shown in Figure 1. I have chosen $\sigma_{max} = 5$ *which leads to* $b = 0.0117$ *and* $\sigma_0 = 0.025$, *compare Appendix 6.1. This leads to* $L_{n,1} = -119.7$. *Neither the optimal solution for* $s = 1$ *nor the one for* $s = 2$ *classifies any point as noise. The right hand side of (4.12) equals* -111.7 *for* $q = 1$ *and* -122.4 *for* $g = 2$. *Thus, the breakdown point is* $\geq \frac{2}{52}$. *Note that the proof of Theorem 4.13 assumes for a possible breakdown of one of the component proportions to 0, that the additional points can be optimally fitted (namely with density value* f_{max}, *therefore the term* $g\log f_{max}$ *in (4.12)) by all components of the optimal solution with* $r < s$, *which is extremely unlikely. Therefore, the values from (4.13) are rather conservative. Better bounds would involve the parameters not only of the optimal solutions for* $r \leq s$, *but also of suboptimal solutions which may fit the additional points better. This may lead to rather complex expressions.*

In this concrete example, the addition of 11 extreme outliers at value 50, say, leads to a breakdown, namely to the classification of one of the two original components as noise, and to the interpretation of the outliers as the second normal component. 10 outliers do not suffice. It is not able to cause a breakdown of one of the normal proportions to 0 by addition of 11 points between the two original components. While important for the proof, the breakdown of a proportion to 0 does not seem to play an important role in practice.

The constellation with a (0,1)-NSD of 45 points combined with a (5,1)-NSD of 5 points leads to the same lower breakdown bound of $\frac{2}{52}$, *but in this case the bound is sharper: 3 outliers at 50 lead to a breakdown of the smaller of the original components; 2 outliers do not suffice. Proportion breakdown again is impossible by addition of so few points, perhaps not even by the addition of arbitrarily many points.*

A more stable data constellation with two clusters is obtained, when a (50,1)-NSD of size 25 is added to the (0,1)-NSD of the same size. In this case (4.12) leads to a minimal breakdown point of $\frac{8}{58}$. *11 outliers at 500 are needed for "empirical" breakdown. The optimal solution for one cluster classifies one of the two NSDs as noise and the other one as the only cluster, while the optimal solution for two cluster again does not classify any point as noise. The same will happen for a larger difference of means, and both sides of (4.12) will stay approximately the same, so that the lower bound for the breakdown point does not increase further. The empirical breakdown behavior does not deviate from the first example.*

4.3 Unknown s

The treatment of s as unknown is favorable for robustness against outliers, because outliers can be fitted by additional mixture components. Generally, for large enough outliers the addition of a new mixture component for each outlier yields a better log-likelihood than any essential change of the original mixture components. That is, gross outliers are almost harmless from the theoretical point of view, except that they let the estimated number of components grow.

Breakdown can occur, however, because added points, usually not outlying, but inside the range of the original data, may lead to a preference of a solution with $r < s$ clusters. (4.14) of Theorem 4.15 gives a necessary condition for the impossibility of breakdown and may serve as a formalization of the "stability" of an s-components solution for a data set in terms of the differences between the optimal log-likelihoods for s and fewer components. However, as in Theorem 4.13, this leads usually to a highly conservative lower bound on the breakdown point.

Breakdown can happen as well because of gross outliers alone, simply because the number of outliers gets so large that the BIC penalty, which depends on n, is increased so much that the whole original data set implodes into fewer than s clusters. The conditions for this are given in (4.16). This cannot happen for the AIC because its penalty does not depend on n.

Again it would be too complex to derive the conditions for the exact breakdown point. An upper and a lower bound for the breakdown point are given here, but they are not very precise.

Theorem 4.15 *Let the assumptions and notations of Theorem 4.13 hold. Let* $\tau_n = (s, \eta_{n,s})$ *be a maximizer of the BIC. If*

$$\min_{r<s} \left[L_{n,s} - L_{n,r} - \frac{1}{2}(5g + 3s - 3r + 2n)\log(n+g) + n \log n \right] > 0, \tag{4.14}$$

then

$$B_n(\tau_n, \mathbf{x}_n) \geq \frac{g}{n+g}. \tag{4.15}$$

If

$$\min_{r<s} \left[L_{n,s} - L_{n,r} - \frac{3}{2}(s-r)\log(n+g) \right] < 0, \tag{4.16}$$

then

$$B_n(\tau_n, \mathbf{x}_n) \leq \frac{g}{n+g}. \tag{4.17}$$

Note that $L_{n,s} - L_{n,r} > \frac{3}{2}(s-r)\log n$ always holds by definition of the BIC. Sufficient conditions for breakdown because of "inliers" depend again on the parameters of certain suboptimal solutions for $r \leq s$ mixture components for $\mathbf{x}_n$ and are presumably too complicated to be of practical use.

Example 4.16 *Consider again the combination of a (0,1)-NSD with 25 points and a (5,1)-NSD with 25 points,* $f = \varphi$ *and* σ_0 *chosen as in Example 4.14. The difference in (4.14) is 3.37 for* $g = 1$ *and* -7.56 *for* $g = 2$*, i.e., the lower breakdown bound is again* $\frac{2}{52}$*. Again, many more points are needed empirically. 13 additional points, equally spaced between 1.8 and 3.2, lead to a final estimation of only one mixture component. It may be possible to find a constellation with fewer points where one component fits better than 2 or more components, but I did not find any. Note that if a breakdown is to be achieved, the additional points between the original clusters are neither allowed to be too widespread, because in this case two components remain optimal, nor too concentrated, because in this case they would generate a third mixture component. Breakdown because of gross outliers according to (4.16) needs more than 650000 additional points!*

The combination of a (0,1)-NSD of 45 points and a (5,1)-NSD of 5 points again leads to the same lower breakdown bound of $\frac{2}{52}$*, and even the empirical robustness of this constellation is* *almost equal to the first: 12 equally spaced points between 1.55 and 3.55 are needed to break down the solution to only one component.*

A mixture of the (0,1)-NSD with 25 points with an (50,1)-NSD of size 25 leads to a lower breakdown bound of $\frac{12}{62}$. *For estimated s, even a breakdown point larger than* $\frac{1}{2}$ *is possible, because new mixture components can be opened for additional points. This may happen empirically already for a mixture of (0,1)-NSD and (50,1)-NSD, because breakdown by addition of gross outliers is impossible unless their number is huge, and breakdown by addition of "inliers" is difficult. For a (0,0.001)-NSD of 25 points and a (100000,0.001)-NSD of 25 points, even the conservative lower breakdown bound is* $\frac{58}{108} > \frac{1}{2}$.

The choice of the t_1*-distribution instead of the Normal leads to a somewhat better breakdown behavior, but the difference is not large: The mixture of a 25 points-(0,1)-NSD and a 25 points-(5,1)-NSD yields a lower breakdown bound of* $\frac{3}{53}$*, and empirically the addition of the 13 inliers mentioned above does not lead to breakdown of one of the two components, but surprisingly to the choice of three mixture components by the BIC. Replacement of the (5,1)-NSD by a (50,1)-NSD again gives a small improvement of the lower bound to* $\frac{13}{63}$.

Remark 4.17 *The possible breakdown point larger than* $\frac{1}{2}$ *here is a consequence of using the addition breakdown definition. A properly defined replacement breakdown point can never be larger than the portion of points in the smallest cluster, because this cluster must be driven to break down if all of its points are suitably replaced. This illustrates that the correspondence between addition and replacement breakdown as established by Zuo (2001) may fail in more complicated setups.*

The addition of a noise component again does not change the breakdown behavior:

Theorem 4.18 *Under* $f_{max} \geq \frac{1}{x_{max,n} - x_{min,n}}$*, Theorem 4.15 holds as well for global maximizers of the BIC, defined so that (2.10) is maximized for every fixed s.*

Example 4.19 *The discussed data examples of two components with 25 points each do not lead to different empirical breakdown behavior with and without estimated noise component, because no point of the original mixture components is classified as noise by the solutions for two Normal components. In the case of the (0,1)-NSD of 45 points and the (5,1)-NSD of 5 points, the solution with one Normal component, classifying the points from the smaller NSD as noise, is better than any solution with two components. That is, no second mixture component exists which could break down.*

Remark 4.20 *While parameter breakdown because of the loss of a mixture component implies classification breakdown of at least one cluster, classification breakdown may occur somewhat earlier than parameter breakdown. Consider again the (0,1)-NSD of 45 points plus the (5,1)-NSD of 5 points. Originally, using simple Normal mixtures, there are two estimated clusters with 45 and 5 points, as expected. The smaller cluster can be broken down by the addition of 6 points, namely 2 points each exactly at the smallest and the two largest points of the (5,1)-NSD. This leads to the estimation of 5 clusters, namely the original (0,1)-NSD, 3 clusters of 3 identical points each, and the remaining 2 points of the (5,1)-NSD. The fifth cluster is most similar to the original one with* $\gamma = \frac{2*2}{2+5} < \frac{2}{3}$*, while no parameter breakdown occurs. For such reasons, an arbitrarily large classification breakdown point is not possible even for very well separated clusters, because not only their separation, but also their size matters. As in Section 4.2, it depends on* σ_0 *how many additional points are needed.*

5 Discussion

It has been shown that none of the discussed mixture model estimators is breakdown robust when the number of components s is assumed as known and fixed. In practice, estimation based on the t-distribution and the addition of a noise component have advantages over the estimation of a simple Normal mixture, because the outliers have to be much larger to break down the estimation. But the number of outliers needed for breakdown stays almost always the same. An alternative is to add an improper uniform distribution with data-independent density as an additional mixture component accounting for noise. Its empirical breakdown characteristics is that the smallest mixture component can be broken down by the addition of outliers of about half of its size. This may be somewhat weaker than trimmed k-means (Garcia-Escudero and Gordaliza 1999), where the breakdown point is related to the size of the smallest component as well, but not worked out exactly up to now.

The more robust way to estimate mixture parameters is the simultaneous estimation of the number of mixture components s by, e.g., the BIC or the AIC. This is almost perfectly breakdown robust against the addition of gross outliers, no matter if mixtures of Normals, t-distributions or Normals with additional noise component are fitted. The only robustness problems arise from the addition of points between the originally estimated mixture components, which may lead to breakdown by estimating a lower number of mixture components. This possibility is extremely data dependent. This kind of breakdown should not be treated as a problem of the methods, but as an internal instability of the dataset with respect to mixture modeling, clustering, respectively. The number of points needed for breakdown according to Theorem 4.15 can be interpreted as a stability characteristics of the dataset. While their precise number is difficult to find, the lower bound (4.15) can be evaluated easily and may serve to compare datasets. However, this has to be interpreted with care because of the conservativeness of the bound.

While including the estimation of s leads to a theoretically satisfactory breakdown behaviour, robustness problems remain in practice, because the global optimum of the loglikelihood must be found. Consider for example a dataset of 1000 points, consisting of 3 well separated clusters of 300 points each, and 100 extremely scattered outliers. The best solution needs 103 clusters. But even for one-dimensional data, the EM-algorithm is very slow for a large number of clusters, and there will be lots of local optima. Therefore, the maximum number of fitted components will often be much smaller than the maximum possible number of outliers, and the use of a proper or improper noise component or t_1-mixtures will be clearly superior to simple Normal mixtures even with estimated s.

I think that it would be a promising area of research to work out classification breakdown points for more general methods of cluster analysis, because such a classification breakdown could provide detailed information about the stability of the clusters, and it would be useful to compare clustering by mixtures with hierarchical methods, say, because they are often used for similar tasks in practice.

I conclude with some comments on the discussed software. Neither `MCLUST` nor `EMMIX` are able to reproduce exactly the results given here. The recent version of `MCLUST` is not able to fit one-dimensional data. Neither `MCLUST` nor `EMMIX` allow the specification of one of a lower scale bound. `MCLUST` produces an error if the EM-iteration leads to a sequence of scale parameters (eigenvalues of the covariance matrix, respectively, in the more than one-dimensional case) converging to 0. This means in particular that no single point can be isolated as its own mixture component, which is crucial for the good breakdown behavior of the methods with estimated s. `EMMIX` terminates the iteration when the loglikelihood seems to converge to infinity. The preliminary iteration results, including one-point-components, are reported, but solutions with variances properly away from 0 are favored. Thus, the current implementations of the Normal mixture estimation with estimated s are essentially non-robust. Addition of a noise component and t-mixtures do better under outliers of moderate size, but they are also not robust against very extreme outliers. The results given here do not favor one of these two approaches over the other, and I think that the implementation of a lower bound for the smallest covariance eigenvalue is the more important issue than the decision between the present implementations.

Note that both packages enable the use of stronger scale restrictions (equivalent to equal variances for all mixture components in the one-dimensional case), which should have roughly the same robustness characteristics for estimated s as the methods treated here. However, such restrictions are often not justified in practice.

6 Appendix

6.1 Choice of the scale restrictions

In most applications, sufficient prior information to specify the scale restriction constants σ_0 of (2.11) and c of (2.12) is not available. A common strategy to avoid a sensible specification of these constants in practice is to compute local maximizers of the log-likelihood from initial values which avoid very small values for the sigmas. This, however, avoids the isolation of single points as clusters, which is crucial for good breakdown behavior for estimated s. The strategy suggested here avoids the necessity of prior information for the specification of c. For the case of σ_0 it will lead to a specification which is clearly interpretable in terms of the subject matter.

Consider s as unknown. A sensible choice of the restriction constants should fulfill two objectives:

1. The constant should be so large that a data subset that looks like a homogeneous cluster is estimated as one component and no single point of it forms a "one-point-component" with a very small scale.
2. The constant should be so small that a gross outlier generates a new component instead of being merged with an otherwise homogeneous data subset.

α-outliers (with $\alpha > 0$ but very small) are defined by Davies and Gather (1993) with respect to an underlying model as points from a region of low density, chosen so that the probability of the occurrence of an outlier is $\leq \alpha$. For a standard Normal distribution, for example the points outside $[\Phi^{-1}(\frac{\alpha}{2}), \Phi^{-1}(1-\frac{\alpha}{2})]$ are the α-outliers. For $\alpha_n = 1-(1-p)^{1/n}$, the probability of the occurrence of at least one α_n-outlier among n i.i.d. points from $\mathcal{N}(0,1)$ is equal to p.

The strategy is as follows: Choose $p = 0.05$, say, and consider the choice of σ_0 for NMML with (2.11) and unknown s. Assume for the moment that at least $n-1$ points come from a $\mathcal{N}(0,1)$ distribution. (Denote $c_0 = \sigma_0$ in this particular setup.) c_0 should be chosen so that it is advantageous to isolate an α_n-outlier as its own cluster, but not a non-outlier. This, of course, depends on the non-outlying data. As "calibration benchmark", form a dataset with n points by adding an α_n-outlier to a (0,1)-NSD (recall Definition 4.10) with $n-1$ points. Choose c_0 so that $C(1) = C(2)$ according to (2.21). This is uniquely possible because $L_{n,1}(\eta_{n,1})$ does not depend on c_0 (as long as c_0 is smaller than the sample variance) and $L_{n,2}(\eta_{n,2})$ increases

n	20	50	100	200	1000
c_0	2.10e-2	4.99e-3	1.66e-3	5.51e-4	4.34e-5
$n_1 = n/2 - 1$	9	24	49	99	499
c_0	2.15e-2	5.25e-3	1.76e-3	5.87e-4	4.57e-5
$n_1 = n/5 - 1$	3	9	19	39	199
c_0	2.25e-2	5.44e-3	1.88e-3	6.35e-4	4.93e-5

Table 1: Minimum scale restriction factor c_0 for NMML with (2.11) and BIC

with decreasing c_0 (because this enlarges the parameter space). For c_0 small enough, the 2-components solution will consist of one component matching approximately the ML-estimator for the NSD, a_2 will approximately equal the outlier and $\sigma_2 = c_0$, so that the decrease of $L_{n,2}(\eta_{n,2})$ gets strict. Resulting values are given in Table 6.1.

The interpretation is as follows: Based on $\sigma_0 = c_0$, a dataset consisting of an $n-1$ point NSD and an α_n-non-outlier will be estimated as homogeneous, while there will be more then one cluster if the nth point is an outlier. It is easily seen that the same will hold for an $n-1$-point (α, σ^2)-NSD and $\sigma_0 = c_0\sigma$. I suggest the use of $\sigma_0 = c_0\sigma_{max}$, where σ^2_{max} is the largest variance such that a data subset with this variance can be considered as "cluster" with respect to the given application. This may not look like an advantage, because the need to specify a lower bound σ_0 is only replaced by the need to specify an upper bound σ_{max}. But the upper bound has a clear interpretation which does not refer to an unknown underlying truth. At least if the mixture model is used as a tool for cluster analysis, points of a cluster should belong together in some sense, and with respect to a particular application, it can usually be said that points above a certain variation can no longer be considered as "belonging together". The estimation of mixture components with larger variance could not be interpreted in this case, even if justified from a purely theoretical point of view.

If $c = c_0$ is chosen with the same strategy applied to NMML with restriction (2.12), c leads to the same behavior for arbitrary (α, σ^2) because of scale equivariance. The resulting values approximately equal the ones given in Table 6.1. (As shown in Lemma 6.1, in this case AIC and BIC will be unbounded for unknown $s \in I\!N$, but this problem can be avoided by specifying an upper bound for s which is smaller than the number of distinct data points.)

A dataset to analyze will usually not have the form "NSD plus outlier", of course. The clusters in the data will usually be smaller than $n-1$ points, and they will have a variance smaller than σ^2_{max}. Assume now that there is a homogeneous data subset of $n_1 < n$ points with variance $\sigma^2 \leq \sigma^2_{max}$. The question arises if an α_{n_1}-outlier, non-outlier, respectively, will be isolated from the cluster in the presence of other clusters elsewhere. c_0 is calculated on the base of the BIC penalty for 1 vs. 2 clusters with n points. That is, the difference in penalty is $3 \log n$. Table 6.1 also gives the c_0-values computed with an NSD of size $n_1 = n/2 - 1$ plus $\alpha_{n/2}$-outlier and of size $n_1 = n/5 - 1$ plus $\alpha_{n/5}$-outlier, but again with penalty difference $3 \log n$ to show which restriction constant would be needed to isolate at least $\alpha_{n/2}$-outliers, $\alpha_{n/5}$-outliers, respectively, from the homogeneous subset of size n_1 under the assumption that the parameters for the rest of the data remain unaffected. The values coincide satisfactorily with the values computed for n, so that these values look reasonable as well for small homogeneous subsets.

With a variance smaller than σ_{max}, an α-outlier with $\alpha > \alpha_n$ is needed to be isolated from a cluster with a variance smaller than σ_{max}, i.e., the broad tendency is that larger components with larger variances are preferred over σ^2_0.

The situation is more complicated for fixed number of components s, because a component can only be added at a particular area of the data if another component vanishes elsewhere. This depends strongly on the constellation of the data. However, the suggestion given here worked reasonable for fixed s as well in some examples.

Note that the theory in Section 4 assumes σ_0, c, respectively, as constant over n, so that it does not apply directly to my suggestion here. However, the differences in log-likelihood caused by the change of σ_0 from n to $n+g$ are expected to be negligible for moderate g and the qualitative breakdown results do not change.

The restriction (2.12) looks more favorable from the point of view of this Section, because the specification of σ_{max} can be avoided by scale equivariance. However, Lemma 6.1 shows that (2.12) does not generalize properly to noise component fitting. To estimate an unknown number of components, an upper bound for s is needed, which is smaller than the number of distinct data points. While such an upper bound is no serious restriction in practice, I expect that (2.12) with unknown s often will prefer badly interpretable solutions with large s and small minimal σ so that s has to be bounded more rigidly.

Lemma 6.1 *The following objective functions are unbounded from above under the restriction (2.12):*

1. *The log-likelihood function (2.10) with fixed s,*
2. *the AIC and BIC of model (2.1) with unknown $s \in I\!N$.*

Proof: Given an arbitrary dataset $x_1, \ldots, x_n$. For (2.10) choose $a_1 = x_1,\ \pi_1 > 0, \sigma_1 \to 0, \pi_0 > 0$. This means that the summand for x_1 converges to ∞ while all others are bounded from below by $\log \frac{\pi_0}{x_{max,n} - x_{min,n}}$. This proves part 1. For part 2 choose $s = n,\ a_1 = x_1, \ldots, a_s = x_n,\ \sigma_1 = \ldots = \sigma_s \to 0$. Thus, $L_{n,s} \to \infty$, and the same holds for AIC and BIC.

6.2 Proofs

Proof of Lemma 2.2:
(2.16) holds because the first sum of (2.14) leads to

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} p_{ij},\ j = 1, \ldots, s$$

(Redner and Walker 1984). (2.17) is simply the separate maximization of the terms of the second sum of (2.14).

For any fixed σ_j^*, the maximizer a_j of (2.17) lies between $x_{max,n}$ and $x_{min,n}$ because of (2.2) and (2.3). Now show that $\sigma_j \leq \frac{\sigma_0 f(0)}{f\left(\frac{x_{max,n} - x_{min,n}}{\sigma_0}\right)}$. By $\sigma_j^* = \sigma_0$,

$$S_j(a_j, \sigma_j) \geq \sum_{i=1}^{n} p_{ij} \log \frac{1}{\sigma_0} f\left(\frac{x_i - a_j}{\sigma_0}\right) \geq n\pi_j \log \frac{1}{\sigma_0} f\left(\frac{x_{max,n} - x_{min,n}}{\sigma_0}\right).$$

For arbitrary σ_j^*,

$$S_j(a_j, \sigma_j^*) \leq n\pi_j(\log f(0) - \log \sigma_j^*).$$

Therefore,

$$\log f(0) - \log \sigma_j \geq \log \frac{1}{\sigma_0} f\left(\frac{x_{max,n} - x_{min,n}}{\sigma_0}\right) \Rightarrow \sigma_j \leq \frac{\sigma_0 f(0)}{f\left(\frac{x_{max,n} - x_{min,n}}{\sigma_0}\right)}$$

as long as $n\pi_j > 0$, proving (2.18).

Proof of Lemma 4.1:
Note first that in case of maximization of (2.10) the density of the noise component $\frac{1}{x_{max,n+g}-x_{min,n+g}}$ converges to 0, so that all arguments, including those used in the proofs of Lemmas 4.2 and 4.3, hold for this case as well.

Assume w.l.o.g. that all $a_{jm},\ j=1,\ldots,s$, lie outside $[x_1-d, x_{n_1}+d]$ for arbitrary $d<\infty$ and large enough m unless $\pi_{jm}\searrow 0$ or $\sigma_{jm}\nearrow\infty$ at least for a subsequence. Consider

$$L_{n,s}(\eta_m, \mathbf{x}_{nm}) = \sum_{i=1}^{n_1}\log\left(\sum_{j=1}^{s}\pi_{jm}f_{a_{jm},\sigma_{jm}}(x_i)\right) + \sum_{i=n_1+1}^{n}\log\left(\sum_{j=1}^{s}\pi_{jm}f_{a_{jm},\sigma_{jm}}(x_i)\right).$$

The first sum converges to $-\infty$ for $m\to\infty$ because of (2.7), and the second sum is bounded from above by $(n-n_1)\log\frac{f(0)}{\sigma_0}$, i.e., $L_{n,s}(\eta_\mu, \mathbf{x}_{nm})\to-\infty$. On the other hand, for $\hat\eta_m$ with $\hat a_{km}=x_{n_k}, \hat\sigma_{km}=\sigma_0,\ \hat\pi_{km}=\frac{1}{h},\ k=1,\ldots,h,$,

$$L_{n,s}(\hat\eta_m, \mathbf{x}_{nm}) \geq \sum_{k=1}^{h} n_k \log\frac{f\left(\frac{x_{n_k m}-x_{(n_{k-1}+1)m}}{\sigma_0}\right)}{h\sigma_0} \geq n\log\frac{f\left(\frac{b}{\sigma_0}\right)}{h\sigma_0} > -\infty.$$

Hence, for large enough m, η_m cannot be ML. Because it should be ML, d must exist so that (4.1) holds for m above some m_0.

Proof of Lemma 4.2:
Proof of (4.3): Suppose that (4.3) does not hold. W.l.o.g. (the order of the a_j does not matter and a suitable subsequence of $(\eta_m)_{m\in\mathbb{N}}$ can be chosen) assume

$$\lim_{m\to\infty}\min\{|x-a_{1m}|:\ x\in\{x_{1m},\ldots,x_{nm}\}\}=\infty.$$

Because of (2.7), $\frac{1}{\sigma_{1m}}f\left(\frac{x_{im}-a_{1m}}{\sigma_{1m}}\right)\to 0\ \forall i$. Because of (2.6) and (4.1),

$$\sum_{j=2}^{s}\pi_{jm}f_{a_{jm},\sigma_{jm}}(x_i) \geq d_{min} = \pi_{min}\frac{1}{\sigma_{max}}f\left(\frac{b+2d}{\sigma_0}\right) > 0,\ i=1,\ldots,n.$$

Thus, for arbitrary small $\epsilon>0$ and m large enough,

$$L_{n,s}(\eta_m, \mathbf{x}_{nm}) \leq \sum_{i=1}^{n}\log\left(\sum_{j=2}^{s}\pi_{jm}f_{a_{jm},\sigma_{jm}}(x_i)\right) + n(\log(d_{min}+\epsilon)-\log d_{min}),$$

and $\log(d_{min}+\epsilon)-\log d_{min}\searrow 0$ for $\epsilon\searrow 0$. Thus, $L_{n,s}$ can be enlarged for small enough ϵ by replacement of $(\pi_{1m}, a_{1m}, \sigma_{1m})$ by $(\pi_{1m}, x_1, \sigma_0)$ in contradiction to η_m being ML.
Proof of (4.4) by analogy to (4.3): Suppose that w.l.o.g. $\sigma_{1m}\to\infty$. Then, $\frac{1}{\sigma_{1m}}f\left(\frac{x_{im}-a_{1m}}{\sigma_{1m}}\right)\to 0\ \forall i$, and replacement of $(\pi_{1m}, a_{1m}, \sigma_{1m})$ by $(\pi_{1m}, x_1, \sigma_0)$ enlarges the log-likelihood.

Proof of Lemma 4.3:
Proof of (4.5): Consider $k\in\{1,\ldots,h\}$. Let $S_k=[x_{(n_{k-1}+1)m}-d, x_{n_k m}+d]$. Because of Lemma 2.2,

$$\sum_{a_{jm}\in S_k}\pi_{jm} = \sum_{a_{jm}\in S_k}\frac{1}{n}\sum_{i=1}^{n}\frac{\pi_{jm}f_{a_{jm},\sigma_{jm}}(x_i)}{\sum_{l=1}^{s}\pi_{lm}f_{a_{lm},\sigma_{lm}}(x_i)}.$$

For $a_{jm} \in S_k$ and $m \to \infty$, $\frac{\pi_{jm} f_{a_{jm},\sigma_{jm}}(x_i)}{\sum_{l=1}^{s} \pi_{lm} f_{a_{lm},\sigma_{lm}}(x_i)} \to 0$ for $i \notin D_k$, while for $i \in D_k$:

$$\left| \frac{\pi_{jm} f_{a_{jm},\sigma_{jm}}(x_i)}{\sum_{l=1}^{s} \pi_{lm} f_{a_{lm},\sigma_{lm}}(x_i)} - \frac{\pi_{jm} f_{a_{jm},\sigma_{jm}}(x_i)}{\sum_{a_{jm} \in S_k} \pi_{lm} f_{a_{lm},\sigma_{lm}}(x_i)} \right| \to 0.$$

This yields $\sum_{a_{jm} \in S_k} \pi_{jm} \to \frac{|D_k|}{n}$, (at least one of the π_{jm} in this sum is bounded away from 0 by (4.1)).
Proof of (4.6): Let $\eta_{kmq} = \arg\max_{\eta} L_{|D_k|,q}(\eta, \mathbf{y}_{km}),\ q \in I\!N$,

$$L_{q_1 \cdots q_h m} = \sum_{k=1}^{h} \left(L_{|D_k|,q_k}(\eta_{kmq_k}) + |D_k| \log \frac{|D_k|}{n} \right).$$

$L_{n,s}(\eta_m) \geq \max_{\sum_{k=1}^{h} q_k = s} L_{q_1 \cdots q_h m}$ can be proven by choice of η according to $\pi_j = \frac{|D_j|}{n} \pi_{jmq_j},\ a_j = a_{jmq_j},\ \sigma_j = \sigma_{jmq_j}, j = 1, \ldots, h$. Further, for m large enough and arbitrarily small $\epsilon > 0$,

$$L_{n,s}(\eta_m) \leq \sum_{k=1}^{h} \sum_{i \in D_k} \log \left(\sum_{a_{jm} \in S_k} \pi_{jm} f_{a_{jm},\sigma_{jm}}(x_i) \right) + \epsilon, \tag{6.1}$$

because for $x_i,\ i \in D_k$, the sum over $a_{jm} \in S_k$ is bounded away from 0 as shown in the proof of Lemma 4.1, while the sum over $a_{jm} \in [x_{(n_{l-1}+1)m} - d, x_{n_l m} + d],\ l \neq k$, vanishes for $m \to \infty$. But

$$\sum_{i \in D_k} \log \left(\sum_{a_{jm} \in S_k} \pi_{jm} f_{a_{jm},\sigma_{jm}}(x_i) \right) - |D_k| \log \left(\sum_{a_{jm} \in S_k} \pi_{jm} \right) \leq L_{|D_k|,q}(\eta_{kmq}),$$

where $q = |\{a_{jm} \in S_k\}|$. Now (4.6) follows from (4.5).

Proof of Theorem 4.4:
Let $\mathbf{x}_{(n+r)m} = (x_1, \ldots, x_n, x_{(n+1)m}, \ldots, x_{(n+r)m}),\ m \in I\!N$, w.l.o.g. $x_1 \leq \ldots \leq x_n,\ x_{(n+k)m} = x_n + km,\ k = 1, \ldots, r$. This fulfills the assumptions of Lemma 4.1 for $h = r + 1$, so that the location parameters for r components must converge to ∞ with $x_{(n+1)m}, \ldots, x_{(n+r)m}$.

Proof of Theorem 4.6:
For a given compact $C(\mathbf{x}_n) \subset I\!R \times I\!R^+$, choose $x_{n+1} > x_{max,n}$ such that $\hat{\theta} \notin C(\mathbf{x}_n)$ for the ML-estimator $\hat{\eta} = \eta_{n+1,1}$, and that for some $\epsilon > 0 : \theta \in C(\mathbf{x}_n) \Rightarrow L_{n+1,1}(\hat{\eta}) > L_{n+1,1}(\eta) + \epsilon$. This is possible because of the well-known non-robustness of $\eta_{n+1,1}$. Add further outliers $x_{n+1+i} = x_{n+1} + im,\ i = 1, \ldots, s-1$. Because of Lemma 4.1, this drives $s-1$ mixture components arbitrarily far away from $\mathbf{x}_{n+1}$, and because of the Lemmas 4.2 and 4.3, the sth mixture component must not fit $\mathbf{x}_{n+1}$ worse by more than ϵ as $\hat{\eta}$ for large enough m, which means that $(a_{sm}, \sigma_{sm}) \notin C$.

Proof of Theorem 4.7: As can be seen from (2.17) and Lemma 2.1, the parameter estimators of $a_j, \sigma_j,\ j = 1, \ldots, s$ can be written as maximizers of a weighted log-likelihood where the weights are given by (2.13). Suppose that there are n original data points $x_1, \ldots, x_n$ and $r < \frac{n}{\nu}$ additional points $x_{n+1}, \ldots, x_{n+r}$. Thus,

$$\frac{\nu}{\nu + 1}(n + r) < n. \tag{6.2}$$

Let $\eta = \eta_{n+r,s}$. Prove by contradiction that there exists $d > 0$, dependent on r but not on $x_{n+1}, \ldots, x_{n+r}$, such that

$$\exists j \in \{1, \ldots, s\} :\ \pi_j \geq d,\ \frac{\sum_{i=n+1}^{n+r} p_{ij}}{\sum_{i=1}^{n+r} p_{ij}} < \frac{1}{\nu + 1} :$$

Suppose that such d does not exist. Then for all $d>0:\ \sum_{i=1}^n p_{ij} < \frac{\nu}{\nu+1}\sum_{i=1}^{n+r} p_{ij}$. Thus,

$$n=\sum_{i=1}^s\sum_{i=1}^n p_{ij} = \sum_{\pi_j<d}\sum_{i=1}^n p_{ij} + \sum_{\pi_j\geq d}\sum_{i=1}^n p_{ij} <$$
$$< |\{j:\ \pi_j<d\}|nd + \tfrac{\nu}{\nu+1}\sum_{\pi_j\geq d}\sum_{i=1}^{n+r} p_{ij} = |\{j:\ \pi_j<d\}|nd + \frac{\nu}{\nu+1}(n+r)\sum_{\pi_j\geq d}\pi_j <$$
$$< snd + \tfrac{\nu}{\nu+1}(n+r) < n$$

for small enough d because of (6.2).

Therefore, there exists at least one mixture component with non-vanishing proportion π_j such that the parameters (a_j,σ_j) are obtained by maximization of a weighted log-likelihood where the weights of the additional points have a proportion smaller than $\frac{1}{\nu+1}$. Get from the proof of Theorem 4.1 of Tyler (1994) that the replacement breakdown point of the ML-estimator for the t_ν-location-scale model is $\geq \frac{1}{\nu+1}$, apart from scale breakdown to 0. The proof holds as well for the maximization of the weighted log-likelihood, where the ratio between the sum of weights of new and original points replaces the ratio between their numbers. According to Zhang and Li (1998, p. 1174), the addition breakdown point is larger or equal than the replacement breakdown point for some suitable dataset with larger n (Zuo 2001 gives an equality result, but the assumptions are not exactly fulfilled here). Therefore, (a_j,σ_j) of the mixture component introduced above must lie in a compact set depending on r and the original data, but not on the added points.

Proof of Theorem 4.8:
The arguments given in the proofs of Theorem 4.4 and Theorem 4.6 apply again, because $\frac{\pi_0}{x_{max,n+r}-x_{\min,n+r}} \to 0$ so that for large enough $x_{(n+r)m}$ the noise component becomes negligible.

Proof of Lemma 4.11:
Proof of (4.9): Recall (2.19). For given x_{n+1}, x_{n+2} can be chosen so large that $p_{(n+1)0}$ gets arbitrarily small for all possible choices of $(a_j,\sigma_j)\in C$ from Lemma 2.2, C determined from $\mathbf{x}_n$ or $\mathbf{x}_{n+1}$. Because of $\sum_{j=0}^s p_{(n+1)j}=1$ and $p_{(n+1)0}<\frac{1}{2}$, say, for all iteration steps apart from the beginning, there must be a $j\in\{1,\ldots,s\}$ such that $p_{(n+1)j}>\frac{1}{2s}$. Because of the non-robustness of the weighted mean, x_{n+1} can be chosen so large that a_j leaves an arbitrary compact set.

Proof of (4.10): Let the parameters before the kth EM-iteration be denoted by $a_j^k,\sigma_j^k,\pi_j^k,p_{ij}^k$. Assume that $x_1\leq\ldots\leq x_n\leq x_{n+1}$ (w.l.o.g.), $p_{(n+1)0}^k>\frac{1}{2}$ and $(a_j^k,\sigma_j^k)\in C$ for $j=1,\ldots,s$. For $k=1$, this holds because of $p_{(n+1)j}^1=0$. Observe $\pi_0^k>\frac{1}{2(n+1)}$ and, for $j=1,\ldots,s$,

$$p_{(n+1)j}^{k+1} = \frac{\pi_j^k\varphi_{a_j^k,\sigma_j^k}(x_{n+1})}{\sum_{j=1}^s\pi_j^k\varphi_{a_j^k,\sigma_j^k}(x_{n+1})+\pi_0^k/(x_{n+1}-x_1)} \leq \frac{1}{\sqrt{2\pi}\sigma_j^k}e^{-\frac{(x_{n+1}-x_n)^2}{2(\sigma_j^k)^2}}2(n+1)(x_{n+1}-x_1).$$

Let

$$\varphi_{min}=\min_{(a,\sigma)\in C,\ x=x_1,\ldots,x_n}\varphi_{a,\sigma}(x).$$

Observe for $j=1,\ldots,s$, where $\pi_j^k\geq\pi_{min}$,

$$p_{1j}^{k+1} = \frac{\pi_j^k\varphi_{a_j^k,\sigma_j^k}(x_1)}{\sum_{j=1}^s\pi_j^k\varphi_{a_j^k,\sigma_j^k}(x_1)+\pi_0^k/(x_{n+1}-x_1)} \geq \pi_{min}\frac{\varphi_{min}}{\varphi_{0,\sigma_0}(0)}$$

Note that a_j^{k+1} is a weighted mean of the weighted mean of x_1 with weight p_{1j}^{k+1} and x_{n+1} with weight $p_{(n+1)j}^{k+1}$ and the corresponding weighted mean of the other points. For x_{n+1} large enough, the weighted mean of x_1 and x_{n+1} is $\leq x_1 + \epsilon$ for arbitrarily small $\epsilon > 0$ because $p_{(n+1)j}^{k+1} \to 0$ for $x_{n+1} \to \infty$. The same argument holds for the scale parameter by use of the weighted mean of $(x_1 - a_j^{k+1})^2$ and $(x_{n+1} - a_j^{k+1})^2$. Because this holds for all iteration steps, this holds as well for every (not necessarily unique) limit point.

Proof of Theorem 4.13:
Let $\mathbf{x}_{n+g} = (x_1, \ldots, x_{n+g})$. Let $\xi^* = \xi_{n+g,s} = \arg\max_{\hat{\xi}} L_{n+g,s}(\hat{\xi}, \mathbf{x}_{n+g})$. For $r < s$,

$$L_{n+g,s} \leq \sum_{i=1}^{n} \log\left(\sum_{j=1}^{r} \pi_j^* f_{\theta_j^*}(x_i) + \sum_{j=r+1}^{s} \pi_j^* f_{\theta_j^*}(x_i) + \pi_0^* b\right) + g \log f_{max}.$$

Assume that the parameter estimators of $s - r$ (i.e., at least one) mixture components leaves a compact set D of the form $D = [\pi_{min}, 1] \times C,\ C \subset I\!R \times I\!R^+$ compact, $\pi_{min} > 0$. Let the mixture components be ordered in such a way that only for $j = 1, \ldots, r < s:\ (\pi_j^*, a_j^*, \sigma_j^*) \in D$. From (2.6): $\sum_{j=1}^{r} \pi_j^* f_{\theta_j^*}(x_i) \geq r\pi_{min} f_{min}$, while $\sum_{j=r+1}^{s} \pi_j^* f_{\theta_j^*}(x_i)$ gets arbitrarily small for large enough D by (2.7). Thus, for arbitrary $\epsilon > 0$ and large enough D:

$$\begin{aligned} L_{n+g,s} &\leq \sum_{i=1}^{n} \log\left(\sum_{j=1}^{r} \pi_j^* f_{\theta_j^*}(x_i) + \pi_0^* b\right) + g \log f_{max} + \epsilon \qquad (6.3) \\ &\leq \max_{r<s} L_{n,r} + g \log f_{max} + \epsilon. \end{aligned}$$

On the other hand, $\hat{\xi}$ could be defined by $\hat{\pi}_0 = \frac{n\pi_0 + g}{n+g},\ \hat{\pi}_j = \frac{n}{n+g}\pi_j,\ \hat{a}_j = a_j,\ \hat{\sigma}_j = \sigma_j,\ j = 1, \ldots, s$. Therefore,

$$\begin{aligned} L_{n+g,s} \geq\ & \sum_{i=1}^{n} \log\left(\sum_{j=1}^{s} \pi_j f_{\theta_j}(x_i) + (\pi_0 + \frac{g}{n})b\right) \\ & + g \log[(\pi_0 + \frac{g}{n})b] + (n+g) \log \frac{n}{n+g} \\ \Rightarrow \quad \max_{r<s} L_{n,r} \geq\ & \sum_{i=1}^{n} \log\left(\sum_{j=1}^{s} \pi_j f_{\theta_j}(x_i) + (\pi_0 + \frac{g}{n})b\right) \\ & + g \log[(\pi_0 + \frac{g}{n})b] + (n+g) \log \frac{n}{n+g} - g \log f_{max} - \epsilon. \end{aligned}$$

This contradicts (4.12) by $\epsilon \to 0$.

Proof of Theorem 4.15:
Add points $x_{n+1}, \ldots, x_{n+g}$ to $\mathbf{x}_n$. Let $C_m(s, \hat{\eta})$ be the value of the BIC for s mixture components and parameter $\hat{\eta}$, applied to the dataset $\mathbf{x}_m,\ m \geq n$. Let $C_m(s)$ be its maximum. With the same arguments as those leading to (6.3), construct for arbitrary $\epsilon > 0$ a suitably large compact $C \subset I\!R \times I\!R^+$, containing the location and scale parameters of all mixture components of $\tau = (s, \eta) = (s, \eta_{n,s})$, and assume that $(a_j^*, \sigma_j^*) \in C$ for only $r < s$ components of $\tau^* = \arg\max_{\hat{s},\hat{\eta}} C_{n+g}(\hat{s}, \hat{\eta})$. Get

$$C_{n+g}(s^*) \leq 2\sum_{i=1}^{n} \log\left(\sum_{j=1}^{r} \pi_j^* f_{\theta_j^*}(x_i)\right) + 2g \log f_{max} + \epsilon - (3s^* - 1)\log(n+g), \qquad (6.4)$$

and, by taking $\hat{s} = s + g,\ \hat{\pi}_j = \frac{n}{n+g}\pi_j,\ j = 1, \ldots, s,\ \hat{\pi}_{s+1} = \ldots = \hat{\pi}_{s+g} = \frac{1}{n+g},\ \hat{\theta}_j = \theta_j,\ j = 1, \ldots, s,\ \hat{a}_{s+k} = x_{n+k},\ \hat{\sigma}_{s+k} = \sigma_0,\ k = 1, \ldots, g,$

$$C_{n+g}(s^*) \geq 2\sum_{i=1}^{n} \log\left(\sum_{j=1}^{s} \frac{n}{n+g}\pi_j f_{\theta_j}(x_i)\right) + 2g\log\frac{f_{max}}{n+g} - (3(s+g)-1)\log(n+g). \tag{6.5}$$

By combination,

$$\begin{aligned}\sum_{i=1}^{n} \log\left(\sum_{j=1}^{s} \pi_j f_{\theta_j}(x_i)\right) - \sum_{i=1}^{n} \log\left(\sum_{j=1}^{r} \frac{\pi_j^*}{\sum_{k=1}^{r}\pi_k^*} f_{\theta_j^*}(x_i)\right) - \epsilon \leq \\ \leq g\log(n+g) - \tfrac{3}{2}(s^* - (s+g))\log(n+g) - n\log\tfrac{n}{n+g} + n\log\left(\sum_{k=1}^{r}\pi_k^*\right) \leq \\ \leq \tfrac{1}{2}(5g + 3s - 3r + 2n)\log(n+g) - n\log n.\end{aligned}$$

This cannot happen for arbitrarily small ϵ under (4.14).

A sufficient condition for breakdown can be derived by explicit contamination. Let $y = x_{n+1} = \ldots = x_{n+g}$. For fixed $\hat{s}$, it follows from Lemma 4.3, that

$$\lim_{y\to\infty} C_{n+g}(\hat{s}) = 2\left(L_{n,\hat{s}-1} + g\log\left(f_{max}\right) + n\log\frac{n}{n+g} + g\log\frac{g}{n+g}\right) - (3\hat{s}-1)\log(n+g).$$

This cannot be maximized by $s^* = \hat{s} > s+1$, because the penalty on s is larger for $n+g$ points than for n points and $s^* - 1$ with parameters maximizing $L_{n,s^*-1}(\hat{\eta}, \mathbf{x}_n)$ must already be a better choice than s for n points unless $s^* \leq s+1$. It follows that the existence of $r < s$ with

$$2L_{n,s} - (3(s+1)-1)\log(n+g) < 2L_{n,r} - (3(r+1)-1)\log(n+g)$$

suffices for breakdown of at least one component, which is equivalent to (4.16).

Proof of Theorem 4.18:
Let $d = \frac{1}{x_{max,n} - x_{min,n}},\ d^* = \frac{1}{x_{max,n+g} - x_{min,n+g}}$. Replace (6.4) by

$$C_{n+g}(s^*) \leq 2\sum_{i=1}^{n} \log\left(\sum_{j=1}^{r} \pi_j^* f_{\theta_j^*}(x_i) + \pi_0^* d^*\right) + 2g\log f_{max} + \epsilon - (3s^*-1)\log(n+g), \tag{6.6}$$

and (6.5) by

$$C_{n+g}(s^*) \geq 2\sum_{i=1}^{n} \log\left(\sum_{j=1}^{s} \frac{n}{n+g}\pi_j f_{\theta_j}(x_i) + \frac{n}{n+g}\pi_0 d\right) + 2g\log\frac{f_{max}}{n+g} - (3(s+g)-1)\log(n+g).$$

(4.15) follows from $d \geq d^*$ in (6.6).

Lemma 4.3 holds as well for maximizers of (2.10), and therefore (4.17) carries over as well.

References

Akaike, H. (1974) A new look at the statistical identification model, *IEEE Transactions on Automatic Control* 19, pp. 716-723

Banfield, J. D. and Raftery, A. E. (1993) Model-Based Gaussian and Non-Gaussian Clustering, *Biometrics* 49, pp. 803-821.

Bozdogan, H. (1994) Mixture Model Cluster Analysis Using Model Selection Criteria and a New Informational Measure of Complexity. In: Bozdogan, H. (ed.), *Multivariate Statistical Modeling, Vol. 2, Proceedings of the First USA/Japan Conference on the Frontiers of Statictical Modeling. An Informational Approach,* Kluwer Academic Publishers, Dordrecht, pp. 69-113.

Bryant, P. and Williamson, A. J. (1986) Maximum likelihood and classification : a comparison of three approaches. In: Gaul, W. and Schader, R. (eds.), *Classification as a Tool of Research,*Elsevier Science, pp. 33-45.

Byers, S. and Raftery, A. E. (1998) Nearest Neighbor Clutter Removal for Estimating Features in Spatial Point Processes, *Journal of the American Statistical Association,* 93, pp. 577-584.

Campbell, N. A. (1984) Mixture models and atypical values, *Mathematical Geology,* 16, pp. 465-477.

Celeux, G. and Soromenho, G. (1996) An entropy criterion for assessing the number of clusters in a mixture, *Journal of Classification,* 13, pp. 195-212.

DasGupta, A. and Raftery, A. E. (1998) Detecting Features in Spatial Point Processes With Clutter via Model-Based Clustering, *Journal of the American Statistical Association,* 93, pp. 294-302.

Davies, P. L. and Gather, U. (1993) The identification of multiple outliers, *Journal of the American Statistical Association,* 88, pp. 782-801.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society B,* 39, pp. 1-38.

DeSarbo, W. S. and Cron, W. L. (1988) A Maximum Likelihood Methodology for Clusterwise Linear Regression, *Journal of Classification,* 5, pp. 249-282.

Donoho, D. L. and Huber, P. J. (1983) The notion of Breakdown point, in Bickel, P. J., Doksum, K. and Hodges jr., J. L. (Eds.): *A Festschrift for Erich L. Lehmann,* Wadsworth, Belmont, CA, pp. 157-184.

Fraley, C., and Raftery, A. E. (1998) How Many Clusters? Which Clustering Method? Answers Via Model Based Cluster Analysis, *Computer Journal*, 41, pp. 578-588.

Gallegos, M. T. (2001) A Robust Method for Clustering Analysis, *NEC Research Index*, `http://citeseer.nj.nec.com/406099.html`.

Garcia-Escudero, L. A., and Gordaliza, A. (1999) Robustness Properties of k Means and Trimmed k Means, *Journal of the American Statistical Association,* 94, pp. 956-969.

Hampel, F. R. (1971) A General Qualitative Definition of Robustness, *Annals of Mathematical Statistics,* 42, pp. 1887-1896.

Hampel, F. R. (1974) The Influence Function and Its Role in Robust Estimation, *Journal of the American Statistical Association,* 69, pp. 383-393.

Hastie, T. and Tibshirani, R. (1996) Discriminant analysis by Gaussian mixtures, *Journal of the Royal Statistical Society B,* 58, pp. 155-176.

Hathaway, R. J. (1985) A constrained formulation of maximum-likelihood estimation for normal mixture distributions, *Annals of Statistics,* 13, pp. 795-800.

Hathaway, R. J. (1986) A Constrained EM Algorithm for Univariate Normal Mixtures, *Journal of Statistical Computation and Simulation,* 23, pp. 211-230.

Huber, P. J. (1964) Robust estimation of a location parameter, *Annals of Mathematical Statistics,* 35, pp. 73-101.

Huber, P. J. (1981) *Robust Statistics,* Wiley, New York.

Keribin, C. (2000) Consistent estimation of the order of a mixture model, *Sankhya A,* 62, pp. 49-66.

Kharin, Y. (1996) *Robustness in Statistical Pattern Recognition*, Kluwer Academic Publishers, Dordrecht.

Lindsay, B. G. (1995) *Mixture Models: Theory, Geometry and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, Hayward.

McLachlan, G. J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture, *Applied Statistics,* 36, pp. 318-324.

McLachlan, G. J. and Basford, K. E. (1988) *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.

McLachlan, G. J. and Peel, D. (2000) *Finite Mixture Models*, Wiley, New York.

Peel, D. and McLachlan, G. J. (2000) Robust mixture modeling using the t distribution, *Statistics and Computing,* 10, pp. 335-344.

Redner, R. A. and Walker, H. F. (1984) Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review,* 26, pp. 195-239.

Rocke, D. M. and Woodruff, D. L. (2000) A Synthesis of Outlier Detection and Cluster Identification. Submitted manuscript, University of California, Davis.

Roeder, K. and Wasserman, L. (1997) Practical Bayesian density estimation using mixtures of normals, *Journal of the American Statistical Association,* 92, pp. 894-902.

Schwarz, G. (1978) Estimating the dimension of a model, *Annals of Statistics* 6, pp. 461-464.

Tyler, D. E. (1994) Finite sample breakdown points of projection based multivariate location and scatter statistics, *Annals of Statistics*, 22, pp. 1024-1044.

Wang, H. H. and Zhang, H. (2002) Model-Based Clustering for Cross-Sectional Time Series Data, *Journal of Agricultural, Biological and Environmental Statistics*, 7, pp. 107-127.

Zhang, J. and Li, G. (1998) Breakdown properties of location M-estimators, *Annals of Statistics,* 26, pp. 1170-1189.

Zuo, Y. (2001) Some quantitative relationships between two types of finite sample breakdown point, *Statistics and Probability Letters* 51, pp. 369-375.